**3GPP TSG-WG SA2 Meeting #149E e-meeting *S2-220xxxx***

**Elbonia, February 14th – 25th, 2022 (revision of S2-220xxxx)**

**Source: Huawei, HiSilicon**

**Title: work assumptions of media traffic characteristics**

**Document for: [Approval]**

**Agenda Item: 9.19**

**Work Item / Release: [FS\_XRM] / Rel-18**

*Abstract: This paper aims to study the traffic characteristics of media service for WT#3.1 and make some consensus for the media traffic characteristics before further technical discussion.*

# 1. Introduction

In FS\_XRM, the WT#3.1 proposes to study the traffic characteristics of media service enabling improved network resource usage and Quek.

Before technical discussions about new key issues and solutions for WT#3, it’s necessary to reach some consensus of the media traffic characteristics within SA2 group.

This paper aims to propose some backgrounds on XR/Media services on codec layer, transporting layer, and to make some work assumption of media traffic characteristics as basis for further discussion.

# 2. Discussion

## 2.1 Real-time and non-real-time XR/media services

**Real-time XR/media services:**

For real-time XR/media services, e.g. cloud gaming, cloud Virtual Reality (VR)/Augmented Reality (AR), online video conference, the media contents are real-time generated/rendered based on user inputs. For such services, strict latency is required. For example, the motion to photo (MTP) time can be 7-15 ms to avoid the user sickness due to the changes of the user viewport as defined in TS 22.261. Usually such services cannot tolerate any packet retransmission between client and server, hence current QoS framework may need to be enhanced to fulfil such service requirements.

**Non-real-time XR/media services:**

For non-real-time XR/media services, such as VoD (Video on-Demand), the media contents are prestored on the server, and the contents can be progressive downloaded (via e.g. MPEG DASH) and buffered on the client before playback. For such non-real-time XR/media services, there is no strict latency requirements.

In this TR, we propose to focus on real-time XR/media services to investigate potential architectural enhancements to support the low latency and high bitrate requirements.

**Proposal 1: In this TR, the XR/media services refers to the real-time media services with low latency requirement, e.g. AR/VR/gaming with cloud rending.**

The following contents in this paper focus on real-time XR/media services.

## 2.2 Traffic characteristics of XR/media service

### 2.2.1 Frame and GOP

A video consists of a sequence of consecutive video frames. Each video frame is a picture encoded/compressed using different codec mechanisms (e.g. H.264/H.265/H.266, AV1 or AVS) for efficient storage and transmission. During encoding, the video spatial and temporal redundancy can be removed using, e.g. inter-frame/intra-frame prediction, motion estimation/compensation. Typically, three major frame/picture types can be used during encoding, i.e. ***I-frame***, ***P-frame*** and ***B-frame***:

* ***I-frame***, as an intra-coded picture, is a complete picture and can be encoded and decoded independently, like a JPG image file.
* ***P-frame***, as a predicted picture, is not a complete frame and only contains the image changes compared to the previous frame. If the reference frame is lost, the P-frame cannot be decoded and displayed.
* ***B-frame***, as a bidirectional predicted picture, contains the changes between the previous and following reference frames. With more reference frames, the compression ratio can be higher. However, the ***B-frame*** can only be decoded when the previous and following reference frames are available.

A ***Group of Pictures (GOP)*** includes a collection of successive video frames. The first frame in a GOP is an I-frame. And the following frames can be P-frames or B-frame.

For example, as shown in figure 1, a GOP contains five frames and the first frame is I frame. The second frame is P frame and refers to the I frame. Similarly, the third frame is P frame, referring to the second frame. The forth frame is a B frame which refers to both the previous and following frames. When delivering the GOP from the application server to a client, the I frame can be directly decoded and displayed. The P frame can be decoded only when the previous frame arrived and is successfully decoded. The B frame can be decoded only when both the previous and following frames arrive and are successfully decoded.



Figure 1 Frame types and GOP

In some codecs, e.g. H.264, ***slice*** can be optionally used during the encoding. A slice is a spatially distinct region of a frame that is encoded separately from any other region in the same frame. Similarly, I-slices, P-slices and B-slices take the place of I, B, P frames.

Frame/slice/GOP are the basic items for encoding/decoding the video. In case a packet of a frame/slice/GOP is lost during the transmission, the frame/slice/GOP will not be fully decoded at the client, which will result in artifacts (e.g. video freezing/blurring/pixelating).

**Proposal 2: Media Unit refers to a group of packets which carry payloads of an element defined in media layer, e.g. frame, GOP, or slice.**

**Proposal 3: A media unit cannot be fully decoded if one packet of the media unit is lost.**

**Proposal 4: Different media units may have different importance.**

### 2.2.2 Frame rate

The ***frame rate*** or frame frequency can be expressed by fps (frame per second), which means the frequency consecutive frames are captured or displayed. Generally, the video sequence can be compressed and encoded following a fixed frame rate (e.g. 30fps, 60fps). Consequently, the output video data traffic also follows a fixed cycle which is also the frame rate. Each burst data corresponds with a single frame (e.g. I, B and P frame). For example, for a 60fps video sequence, there are 60 bursts per second and the interval of two consecutive bursts is 1/60 second (16.67ms).

**Proposal 5: The traffic of a video stream composes of periodic bursts with periodicity of 1/(frame rate).**

# 4. Proposal(s)

It is proposed to capture the proposals as work assumption in TR 23.700-60 and add clause 2.2 in discussion part in the TR as informative annex.

\* \* \* \* First change (based on skeleton)\* \* \* \*

# 4 Architectural Assumptions and Principles

Editor’s Note: This clause will document any architectural assumptions and principles.

1. In this TR, 5GS QoS framework defined in Rel-17 is used as basis for further potential enhancement to support more efficient XR/media services.
2. In this TR, the XR/media services refer to the interactive media services with low latency requirement, e.g. AR/VR/gaming with cloud rending.
3. Media unit refers to a group of packets which carry payloads of an element defined in media layer, e.g. frame, GOP, or video slice.
4. A media unit cannot be fully decoded by the client if one packet of the media unit is lost.
5. Different media units may have different importance.
6. The traffic of a video stream composes of periodic bursts with a periodicity of 1/(frame rate).

\* \* \* \* Second change (all new text) \* \* \* \*

# Annex X (informative) Traffic characteristics of real-time XR/media service

## X.1 Frame and GOP

A video consists of a sequence of consecutive video frames. Each video frame is a picture encoded/compressed using different codec mechanisms (e.g. H.264/H.265/H.266, AV1 or AVS) for efficient storage and transmission. During encoding, the video spatial and temporal redundancy can be removed using, e.g. inter-frame/intra-frame prediction, motion estimation/compensation. Typically, three major frame/picture types can be used during encoding, i.e. ***I-frame***, ***P-frame*** and ***B-frame***:

* ***I-frame***, as an intra-coded picture, is a complete picture and can be encoded and decoded independently, like a JPG image file.
* ***P-frame***, as a predicted picture, is not a complete frame and only contains the image changes compared to the previous frame. If the reference frame is lost, the P-frame cannot be decoded and displayed.
* ***B-frame***, as a bidirectional predicted picture, contains the changes between the previous and following reference frames. With more reference frames, the compression ratio can be higher. However, the ***B-frame*** can only be decoded when the previous and following reference frames are available.

A ***Group of Pictures (GOP)*** includes a collection of successive video frames. The first frame in a GOP is an I-frame. And the following frames can be P-frames or B-frame.

For example, as shown in figure 1, a GOP contains five frames and the first frame is I frame. The second frame is P frame and refers to the I frame. Similarly, the third frame is P frame, referring to the second frame. The forth frame is a B frame which refers to both the previous and following frames. When delivering the GOP from the application server to a client, the I frame can be directly decoded and displayed. The P frame can be decoded only when the previous frame arrived and is successfully decoded. The B frame can be decoded only when both the previous and following frames arrive and are successfully decoded.



Figure 1 Frame types and GOP

In some codecs, e.g. H.264, ***slice*** can be optionally used during the encoding. A slice is a spatially distinct region of a frame that is encoded separately from any other region in the same frame. Similarly, I-slices, P-slices and B-slices take the place of I, B, P frames.

Frame/slice/GOP are the basic items for encoding/decoding the video. In case a packet of a frame/slice/GOP is lost during the transmission, the frame/slice/GOP will not be fully decoded at the client, which will result in artifacts (e.g. video freezing/blurring/pixelating).

## X.2 Frame rate

The ***frame rate*** or frame frequency can be expressed by fps (frame per second), which means the frequency consecutive frames are captured or displayed. Generally, the video sequence can be compressed and encoded following a fixed frame rate (e.g. 30fps, 60fps). Consequently, the output video data traffic also follows a fixed cycle which is also the frame rate. Each burst data corresponds with a single frame (e.g. I, B and P frame). For example, for a 60fps video sequence, there are 60 bursts per second and the interval of two consecutive bursts is 1/60 second (16.67ms).

\* \* \* \* End of Changes \* \* \* \*